Check for updates

Journal of Korea Robotics Society (2024) 19(3):293-310
https://doi.org/10.7746/jkros.2024.19.3.293 ISSN: 1975-6291 / eISSN: 2287-3961

언어-기반 제로-샷 물체 목표 탐색 이동 작업들을 위한 인공지능 기저 모델들의 활용

Utilizing AI Foundation Models for Language-Driven Zero-Shot Object Navigation Tasks

최 정 현[1] · 백 호 준[1] · 박 찬 솔[1] · 김 인 철[†]
Jeong-Hyun Choi[1], Ho-Jun Baek[1], Chan-Sol Park[1], Incheol Kim[†]

Abstract: In this paper, we propose an agent model for Language-Driven Zero-Shot Object Navigation (L-ZSON) tasks, which takes in a freeform language description of an unseen target object and navigates to find out the target object in an inexperienced environment. In general, an L-ZSON agent should able to visually ground the target object by understanding the freeform language description of it and recognizing the corresponding visual object in camera images. Moreover, the L-ZSON agent should be also able to build a rich spatial context map over the unknown environment and decide efficient exploration actions based on the map until the target object is present in the field of view. To address these challenging issues, we proposes AML (Agent Model for L-ZSON), a novel L-ZSON agent model to make effective use of AI foundation models such as Large Language Model (LLM) and Vision-Language model (VLM). In order to tackle the visual grounding issue of the target object description, our agent model employs GLEE, a VLM pretrained for locating and identifying arbitrary objects in images and videos in the open world scenario. To meet the exploration policy issue, the proposed agent model leverages the commonsense knowledge of LLM to make sequential navigational decisions. By conducting various quantitative and qualitative experiments with RoboTHOR, the 3D simulation platform and PASTURE, the L-ZSON benchmark dataset, we show the superior performance of the proposed agent model.

Keywords: Language-driven Zero-shot Object Goal Navigation, Goal Grounding, Spatial Context Mapping, LLM Prompting, Exploratory Action Decision

1. 서 론

최근들어 대규모 언어 모델(Large Language Model, LLM)과 시각-언어 모델(Vision-Language Model, VLM) 등 인공지능 기저 모델(AI foundation model)들의 발전이 거듭되고 있는 가운데, 체화 인공지능(embodied AI) 및 로보틱스(robotics) 분야에서도 인공지능 기저 모델들을 활용하여 새로운 환경에서 보다 유연하고 지능적으로 동작할 수 있는 에이전트들이 활발히 개발되고 있는 추세이다. 본 논문에서는 언어-기반 제로-샷 물체 목표 탐색 이동(Language-driven Zero-Shot Object Navigation, L-ZSON) 작업을 위한 새로운 에이전트 모델을 제시한다. L-ZSON은 로봇 에이전트가 이전에 경험한 적이 없는 새로운 미지의 환경(unseen environment)에서 목표 물체에 관한 자연어 묘사 (natural language description of the target object)를 토대로, 실시간 입력 영상만을 이용하여 해당 목표 물체를 찾아 이동 행동을 계획하고 실행하는 작업을 가리킨다. 전통적인 물체 목표 탐색 이동(Object Goal Navigation, OGN) 작업에서는 미지의 환경에 놓여진 에이전트에게 미리 정해진 범주(category)의 한 물체를 찾아 이동하도록 요구한다. 반면에, 제로-샷 물체 목표 탐색 이동(Zero-Shot Object Navigation, ZSON) 작업은 해당 환경을 이용한 어떠한 사전 훈련(training) 과정없이도 목표

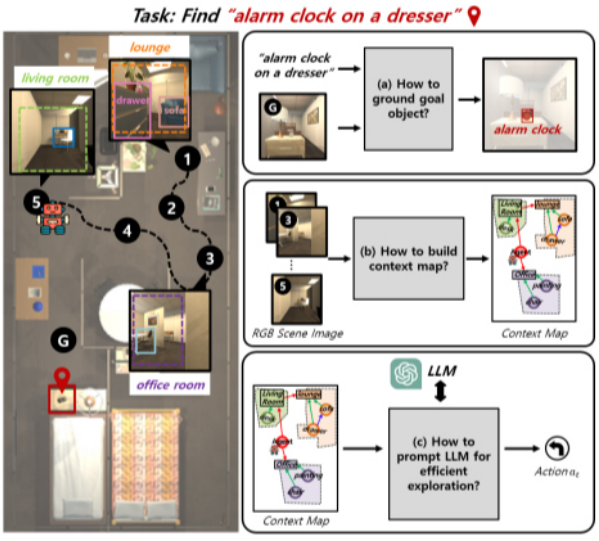
Received : May. 20. 2024; Revised : Jul. 16. 2024; Accepted : Aug. 16. 2024
1. Student, Computer Science, Kyonggi University, Suwon, Korea (wjdgus21@kyonggi.ac.kr, qorgh3466@gmail.com, json@kyonggi.ac.kr)
† Full Professor, Corresponding author: Computer Science, Kyonggi University, Suwon, Korea (kic@kyonggi.ac.kr)

Copyright©KROS

물체를 식별해내도록 전통적인 물체 목표 탐색 이동(OGN) 작업을 확장한 것이다. 여기에 언어-기반 제로-샷 물체 목표 탐색 이동(L-ZSON) 작업은 자유 형식의 자연어 목표 물체 묘사(freeform natural language description of the target object)로부터 해당 목표 물체를 식별해내도록 제로-샷 물체 목표 탐색 이동(ZSON) 작업의 난이도를 한 단계 더 높인 작업이다.

[Fig. 1]은 L-ZSON 작업의 한 예와 더불어 L-ZSON 작업을 효과적으로 수행할 수 있는 에이전트 모델의 설계 시에 고려해야 할 도전적인 이슈(challenging issue)들을 나타낸다. 이 L-ZSON 작업에서는 에이전트가 여러 방들과 가구들로 구성된 미지의 한 가정환경에서 자연어 묘사인 “alarm clock on a dresser”를 토대로 해당 목표 물체를 찾아가야 한다. 이때, 에이전트는 매 순간 내장 카메라를 통해 실시간 RGB-D 영상을 입력받고, 6가지 이산화된 이동 행동들인(navigation actions) {MoveForward, TurnLeft, TurnRight, LookUp, LookDown, Stop} 중 하나를 선택해 실행하는 것으로 가정한다. 그리고 에이전트가 적절한 이동 행동들을 통해 제한 시간 내에 목표 물체인 alarm clock에 접근할 수 있으면, 해당 L-ZSON 작업을 성공적으로 완수한 것으로 간주한다.

L-ZSON 작업을 위한 에이전트 모델을 설계할 때 고려해야 할 중요한 도전적 이슈 중 하나는 [Fig. 1]의 우측 (a)에 표시한 것처럼, “alarm clock on a dresser”와 같이 목표 물체의 속성 혹은 다른 물체와의 공간적 관계를 설명하는 자연어 묘사가 주어졌을 때, 어떻게 실시간 RGB-D 입력 영상에서 해당 목표 물체를 정확히 식별해내는가 하는 목표 물체 그라운딩(goal grounding) 문제이다. 기존의 L-ZSON 모델들[1-3]에서는 대부분 사전 학습된 제로-샷 개방형 어휘(pre-trained zero-shot open-vocabulary) 물체 탐지기들인 OWL-ViT[4], GLIP[5] 등을 활용하여 목표 물체를 시각적으로 그라운딩(visually grounding) 하고자 하였다. 이러한 제로-샷 물체 탐지 신경망 모델들은 학습 단계에서는 주어지지 않은 범주 레이블(class label)을 가진 일부 물체들도 추가로 인식 가능하다는 유연성은 있다. 하지만 이들을 이용할 경우 한 두 단어(word)로 구성된 물체의 범주 레이블(class label)로 목표 물체 묘사가 제한될 뿐만 아니라 , 아직은 이 제로-샷 물체 탐지기들의 인식 성능도 낮다는 문제점이 있다. 따라서 이들은 L-ZSON 작업들에서 요구하는 것처럼 색상(color), 모양(shape), 재질(material) 등 다양한 목표 물체의 속성(attribute)들과 함께 다른 물체들과의 공간적 관계(spatial relationship)들을 설명하는 자유 형식의 자연어 묘사를 심층적으로 이해하고 해당 목표 물체를 실시간 RGB-D 입력 영상에서 탐지해내는데는 한계가 있다.



[Fig. 1] Issues of designing L-ZSON agent model

L-ZSON 에이전트 설계의 또 다른 중요한 이슈는 [Fig. 1]의 우측 (b)에 표시한 것처럼, 실시간 RGB-D 입력 영상들로부터 에이전트가 미지의 환경을 심층적으로 이해할 수 있는 정확한 공간적 맥락 지도(spatial context map)를 어떻게 생성할 것이냐 하는 문제이다. [Fig. 1]의 예에서 보듯이, 에이전트가 L-ZSON 작업을 효과적으로 수행하기 위해서는 해당 가정환경에서 장애물들이 점유하고 있는 지역과 에이전트가 자유롭게 이동 가능한 지역을 구분할 수 있는 것은 물론이고, 나아가 해당 환경을 구성하는 방들과 가구를 포함한 기타 물체들, 그리고 그들 간의 공간적 위치 관계들까지 맥락 지도에 정확히 표현되는 것이 바람직할 것이다. 기존의 L-ZSON 모델들[1-3]은 대부분 이동 계획 수립에 꼭 필요한 장애물 점유 지도(occupancy map)만을 생성해서 이용하거나 또는 시각적으로 인식된 주변 물체들을 중심으로 환경에 대한 의미적 지도(semantic map)를 작성하여 이용하는 경우도 있었다. 하지만 낮은 물체 인식률로 인해 의미적 지도의 정확도도 낮고, 지도에 각 개별 물체들의 위치가 표시되기는 하지만 방과 방, 방과 물체, 물체와 물체들 간의 공간적 위치 관계들은 명시적으로 지도에 반영되지 않는다는 한계성이 존재한다.

L-ZSON 에이전트 설계의 중요한 마지막 이슈는 목표 물체가 에이전트의 시야에 포착될 때까지, 어떻게 효율적으로 탐험적 행동들(explorative actions)을 수행하도록 할 것인가에 대한 문제이다. 많은 경우 L-ZSON 에이전트들은 작업의 대부분 시간을 목표 물체를 시각적으로 찾을 때까지 환경 공간을 탐험하는데 소모한다. 따라서 이러한 에이전트의 탐험적 행동 정책(exploration policy)은 L-ZSON 작업의 효율성(efficiency)을 좌우하는 매우 중요한 요소가 될 수 있다[6]. 일반적으로 시야에 있지 않은 목표 물체를 찾아가려면, 목표 물체가 위치해 있을 장소를 짐작할 수 있는 기본적인 상식(commonsense knowledge)이나 경험(experience)에 의존할 수 밖에 없다. 예컨대, 침대는 침실에 있을 가능성이 높고, 소파는 거실, TV는 거실이나 침실

에 있을 가능성이 높다는 등의 상식을 이용해, 해당 목표 물체가 놓여 있을 가능성이 높은 장소들을 우선적으로 탐험해보는 행동 정책은 목표 물체 발견에 큰 도움이 될 것이다. 최근에는 대용량의 문서 집합들로 사전 학습된 GPT-3[7], LLaMA2[8] 등과 같은 대규모 언어 모델(Large Language Model, LLM)이 크게 발전함에 따라, 이 언어 모델들이 가진 자연어 대화 능력 이외에 이 모델들이 wikipedia 와 같은 문서 집합들로부터 추출해서 내재하고 있을 다양한 상식들을 활용해보고자 시도하는 연구들도 등장하게 되었다[1-3]. 하지만 대부분의 기존 연구들에서는 (1) LLM에게 어떤 맥락정보를 제공하고, 이로부터 어떤 응답을 받아낼 것인지 구체적인 프롬프팅(prompting) 방식과 (2) LLM의 응답을 탐험적 행동 결정에 구체적으로 어떻게 활용할 것인지 등 중요한 세부 이슈들을 자세히 다루지는 않았다.

본 논문에서는 위에서 언급한 각 설계 이슈별 기존의 연구들의 한계성을 극복하기 위해, 새로운 L-ZSON 에이전트 모델인 AML (Agent Model for L-ZSON)을 제안한다. 제안 모델 AML은 자연어 목표 물체 묘사의 시각적 그라운딩, 즉 목표 물체 그라운딩(goal grounding) 이슈와 관련해서는 참조 영상 분할(Referring Image Segmentation, RIS) 이 가능한 인공지능 기저 모델의 하나인 GLEE[9]를 채용한다. GLEE는 자유 형식의 자언어 묘사에 부합되는 특정 물체를 입력 영상 안에서 찾아줄 수 있도록 사전 학습된 시각-언어 모델(Vison-Language Model, VLM)이다. 그러나 GLEE 모델은 참조 영상 분할 모델의 특성상 참조 대상인 목표 물체가 영상 내에서 관측되지 않을 때에도 영상 안의 다른 물체를 목표 물체로 식별해내는 약점을 가지고 있다. 이러한 특성은 작업 대부분의 시간 동안 목표 물체가 시야 내에 들어오지 않는 LZSON 작업에서는 치명적인 문제가 될 수 있다. 이를 보완하기 위해 제안 모델에서는 목표 물체 그라운딩을 위해 GLEE를 활용하되, 또 다른 시각 인식 기저 모델인 GLIP을 병행 사용하여 GLEE의 식별 능력을 효과적으로 보완한다. 또 제안 모델 AML은 공간적 맥락 지도 생성(spatial context map building) 이슈와 관련해서는 시각 인식 기저 모델(Vision Model, VM)인 제로-샷 물체 탐지기(zero-shot object detector) GLIP[5]을 이용하여 이동 계획 수립에 필요한 장애물들의 점유 지도뿐만 아니라 환경을 구성하고 있는 방, 가구, 기타 물체들 간의 공간적 배치 관계들도 나타낼 수 있는 의미적 맥락 지도도 생성한다. 또한 탐험적 행동 정책(exploration policy) 이슈와 관련해서는, 제안 모델 AML은 LLM의 상식을 효과적으로 이용하기 위한 맥락정보 기반의 프롬프팅(context-based prompting) 방법을 제시할 뿐만 아니라, LLM의 응답을 토대로 구체적인 탐험적 행동 결정을 위한 확률적 추론(probabilistic reasoning) 방법도 제시한다.

본 논문에서는 가상 실내 환경들을 제공하는 3차원 시뮬레이터인 RoboTHOR[10]와 L-ZSON 벤치마크 데이터 집합인 PASTURE[1]를 이용한 다양한 정량적, 정성적 실험들을 수행함으로써, 제안 모델의 우수성을 입증한다. 본 논문의 2장에서는 관련 선행 연구들을 소개하고, 3장에서는 제안 모델의 설계에 대해 자세히 설명한다. 4장에서는 제안 모델의 구현과 실험에 관해 소개하고, 5장에서는 결론과 향후 연구를 정리한다.

2. 관련 연구

2.1 목표 물체 그라운딩

목표 물체 그라운딩은 에이전트가 목표 물체를 지칭하는 자연어 문장 표현을 개념적으로 이해하고, 입력 영상 속 해당 물체의 정확한 위치를 특정해내는 과정이다. 특히 L-ZSON 작업에서는 다양한 속성을 묘사한 문장을 제시하므로 시각적 특징뿐만 아니라 공간적 관계, 상황 맥락 등을 함께 이해하는 능력이 요구된다. 이를 위해 기존의 L-ZSON 모델[1-3]들은 텍스트와 입력 영상을 동시에 처리하는 제로-샷 물체 탐지 모델(zero-shot object detector)[4,5]을 채용하였다. 이로써 미지의 환경 내 새로운 범주(category)의 물체가 존재하는 장면을 설명한 텍스트 프롬프트(text prompt)로부터 영상 속 물체의 위치를 예측해낼 수 있게 되었다. 이는 YOLO[11]나 DETR[12]과 같이 별도의 언어적 이해 없이 사전 훈련된 범주 레이블만 탐지하는 기존의 물체 탐지 모델들의 접근 방식 대비, 유연하며 포괄적인 물체 탐지를 가능하게 한다는 장점을 보인다.

CoW[1] 모델은 OWL-ViT[4]를 활용하여 목표 물체를 그라운딩하고자 하였다. OWL-ViT는 비전 트랜스포머(vision transformer)[13] 기반의 영상 인코더(image encoder)와 트랜스포머 기반의 텍스트 인코더를 이용하여 입력 RGB 영상과 텍스트 형태의 목표를 각각 시각적, 언어적 임베딩 벡터로 생성하고, 의미적인 연결을 통해 영상 내에서 위치를 특정해낸다. 이를 위해 사전 훈련 과정에서 MS COCO와 Open-Image 데이터 집합에 포함된 대규모의 영상-텍스트 쌍을 활용하여, 영상 내 물체와 관련된 텍스트의 의미적 맥락을 매핑(mapping)하는 방식으로 학습한다. 이로써 새로운 물체 범주도 그라운딩 가능하지만, 사전 훈련 데이터 집합의 크기와 다양성이 제한적이므로 다양한 시나리오나 물체 범주에 대한 그라운딩 작업에서 제한적인 성능을 보인다.

이에 반해 LGX[2]와 ESC[3] 모델에서는 스윈 트랜스포머(swin transformer)[14] 기반의 영상 인코더와 트랜스포머 기반의 텍스트 인코더를 결합한 구조를 가지는 GLIP[5] 모델을 활용하였다. 이때 두 모델에서 채용한 GLIP-L은 FourODS, GoldG, Cap4M 등 3가지 대규모 영상-텍스트 데이터 집합에서 사전 훈련되었다. 또, 물체 탐지와 그라운딩 작업을 분리하여 각각 사전 훈련을 수행하는 OWL-ViT와 달리 물체 탐지와 그라운딩 과정을 통합

하여 사전 훈련을 수행하므로 높은 모델 범용성을 가지며 영상과 텍스트 간의 상호작용을 더 정확하게 이해하는 강력한 그라운딩 성능을 보인다. ZSON 작업의 목표 물체 그라운딩에 적용하였을 때도 OWL-ViT 대비 더 높은 작업 성능을 보여주며 그 효과를 증명하였다[3]. 그러나 L-ZSON의 목표 물체의 여러 속성들을 포함하는 자연어 표현식에 가장 적합한 영역을 찾아내는 참조 표현 이해(Referring Expression Comprehension, REC) 기능은 제공하지 못하여 L-ZSON의 목표 물체 그라운딩에 적용하기에는 제한적이다.

2.2 전역적 맥락 지도

L-ZSON 작업에서 에이전트가 전역적 맥락 지도(global context map)를 통해 환경 및 현재 작업 상황을 인식하는 것은 매우 중요하다. 전통적으로 로봇 또는 에이전트 분야에서 이동 작업을 위해 깊이 영상을 기반으로 그리드(grid) 내 각 셀(cell) 혹은 그래프(graph) 내 각 노드의 점유 여부를 나타내는 점유 지도를 사용하였다. 특히, 시각적 탐색 이동 작업에서 휴리스틱 탐색(heuristic search) 알고리즘 중 하나인 A* 또는 빠른 전진 방식(Fast Marching Method, FMM) 등 경로 계획 알고리즘과 결합되어 주로 사용되었다[1,2,6]. Cow[1], LGX[2]에서는 각 셀을 자유(free), 점유(occupied), 미탐험(unexplored) 영역의 3가지로 구분하는 그리드 형태의 점유 지도를 사용하였다. VoroNav[6]에서는 각 노드를 에이전트와 직접 연결된 이웃 노드(neighbor nodes), 미탐험 영역에 인접한 경계를 의미하는 탐험 노드(exploratory nodes), 위의 두 범주에 속하지 않는 모든 다른 노드인 일반 노드(ordinary nodes)의 3종류로 구분하는 그래프 형태의 점유 지도를 사용하였다. 이때, 노드를 연결하는 간선은 인접 노드 간의 이동 가능성을 나타내며, 이를 토대로 경로를 계획한다. 이러한 점유 지도는 에이전트 위치와 이동 가능 영역 정보를 제공하지만, 작업 환경을 구성하는 물체 간의 상호작용 정보를 표현하기는 어렵다. 반면, ESC[3] 모델에서는 제로-샷 물체 탐지기인 GLIP[5]을 이용하여 각 셀에 물체들과 방마다 탐지 신뢰도를 저장한 의미적 지도를 생성함으로써 작업 환경의 공간 구성을 집중적으로 파악하려고 하였다. 하지만 작업 수행 환경의 방과 물체 간의 공간 관계, 물체와 물체 간의 공간 관계까지 충분히 이해한다고 보기 어렵다.

한편, 물체 목표 탐색 이동(OGN) 작업에서 환경 물체와 방 사이 공간 관계를 집중적으로 파악하고자 한 연구들도 있었다. ORG[15] 모델은 관측한 물체들의 범주와 위치 기반의 관계를 표현하는 물체 관계 그래프(object relation graph)를 통해 목표 물체가 시야 내에 존재하지 않을 때, 목표와 관련된 물체들을 탐험 행동에 활용하고자 하였다. 하지만 L-ZSON 작업 환경에서는 물체의 종류와 놓여있는 위치, 그리고 그들이 존재하는 방 등이 일관성이 부족하여 학습을 통해서는 공간 구성에 대해 충분히 이해한다고 보기는 어렵다. 이를 보완하기 위해 HOZ[16] 모델은 각 노드를 장면 노드(scene node), 영역 노드(zone node), 그리고 물체 노드(object node)로 구성하고, 간선은 인접 확률을 나타내는 그래프 형태의 지도를 구성하였다. 이러한 구조는 작업 환경을 계층적으로 구조화하고 작업 환경을 영역 단위로 패턴화하여 규칙성을 활용하고자 한 시도이지만, 간선이 단순히 인접성만을 표현하므로 구체적인 영역(방)과 물체 간의 관계 종류를 알기는 어렵다는 한계를 보인다.

2.3 언어 모델 기반 탐험적 행동 결정

L-ZSON 작업에서는 대부분 목표 물체가 시야 범위 내 들어오지 않기 때문에 목표 물체를 효율적으로 찾기 위한 탐험적 행동을 결정하는 것이 필수적이다. 최근 GPT-3[7], LLaMA[8]와 같은 대규모 언어 모델(LLM)의 발전에 힘입어 L-ZSON 모델들[1-3]은 사전-학습된 LLM의 기본적인 상식(commonsense knowledge)을 활용하여 탐험적 행동에 필요한 추가적인 단서를 얻어내고자 하였다. 이 과정에서 (1) LLM에게 어떤 시-공간적 맥락정보를 제공하고, 이로부터 어떤 응답을 받아낼 것인지 제약을 주는 프롬프팅(prompting) 방식 결정 문제와 (2) 효과적인 탐험적 행동 결정을 위해 LLM의 응답을 어떻게 활용할 것인가의 두 가지 세부 이슈로 나눌 수 있다.

먼저 프롬프팅 방식 결정 문제와 관련한 기존 연구들을 살펴보면, CoW[1] 모델은 ‘wooden toy airplan’과 같은 목표 물체에 대한 언어적 묘사를 제공하여, 목표가 일반적으로 어느 방에 위치하는가에 대한 상식적인 추론을 수행하였다. 하지만 목표 물체와 관련된 상식에만 의존하므로 LLM이 작업 환경에 존재하지 않는 방을 제시하거나 에이전트의 위치를 고려하지 않은 답변을 제공하는 문제가 발생할 수 있다. LGX[2] 모델은 에이전트의 주변의 물체들의 범주 레이블에 대한 지역적 맥락정보를 추가 제공하여, 에이전트가 목표 물체에 도달하기 위해 어떤 주변 물체들을 활용해야 하는지에 대한 상식적인 추론을 수행하도록 하였다. 그러나 현재 위치에 기반한 주변 환경만 참조하므로 공간 구조를 충분히 반영하지 못할 수 있다. 또한, 에이전트의 현재 위치에 기반한 맥락정보에 제한되므로 지나온 경로 상의 중요한 맥락정보를 반영하기는 어렵다는 한계를 보인다.

반면, ESC[3] 모델은 작업 환경에 포함된 모든 방들과 탐험 과정에서 탐지된 환경 물체들의 범주를 포함한 전역적 맥락정보를 추가로 활용하였다. 해당 모델에서는 에이전트의 현재 시야에 탐지된 환경 물체들과 목표 물체 간의 연관성을 개별적으로 평가하여, 목표 물체를 찾을 수 있는 잠재적 위치에 대한 상식적인 추론을 수행하도록 하였다. 하지만 물체와 방 사이의 공

간적 관계나 방과 방 사이의 연결 관계를 충분히 고려하지 못하는 한계점이 존재한다.

한편, 탐험을 위한 LLM의 응답 활용 문제와 관련한 기존 연구들을 살펴보면, CoW[1], LGX[2] 모델에서는 LLM으로부터 목표 물체가 자주 발견되는 특정 방이나 물체를 식별하기 위한 지시어를 답변받고, 해당 방 또는 물체까지의 최단 경로로 이동함으로써, LLM의 추론 결과를 바로 이동 행동으로 적용하고자 하였다. 하지만 일반적인 상식에만 의존하는 방법이 에이전트 작업 환경의 다양성과 복잡성을 충분히 반영하지 못한다는 한계가 있다. 이와 달리, ESC[3] 모델에서는 LLM으로부터 받은 목표 물체와 환경 물체, 그리고 방과 방 사이의 동시 출현(co-occurrence) 확률을 활용하여, 확률적 소프트 논리(Probabilistic Soft Logic, PSL)를 기반으로 다음 탐험 지점을 결정한다. 이 방식은 LLM에 대한 과도한 의존을 줄이고, 목표 물체가 나타날 가능성이 높은 환경 물체나 방의 경계를 우선적으로 선택하는 탐험적 행동을 수행하도록 하였다. 하지만, 첫 번째 프롬프팅 단계에서 응답받는 방과 두 번째 프롬프팅 단계에서 응답받는 물체들 간의 직접적인 연관성이 부족하다는 한계점이 존재한다.

3. L-ZSON 에이전트 모델 설계

3.1 L-ZSON 작업 정의

언어-기반 제로-샷 물체 목표 탐색 이동(L-ZSON) 작업에서 에이전트는 미지의 환경에서 자유 형식의 자연어 묘사로 주어지는 목표 물체를 찾아가야 한다. 즉, 에이전트가 한 번도 경험한 적 없는 새로운 환경 집합$\{E_{new}\}$와 새로운 목표 물체 집합$\{g_{new}\}$이 주어졌을 때, 목표 물체 범주와 관련하여 훈련된 경험 없이도$\{g_{new}\}$을 포함하는$\{E_{new}\}$에서 탐색 이동 작업을 수행해야 한다. 에피소드 시작 시, 에이전트는 작업 환경 $E \in E_{new}$에서의 초기 위치 P_0에서부터 주어진 자연어 문장 $sentence(g)$에서 설명된 목표 물체 $g \in g_{new}$를 찾아가야 한다. 각 시간 단계마다 에이전트는 입력 센서 정보 O_t를 기반으로 하나의 행동 a_t을 선택한다. O_t는 672×672 크기의 에이전트 중심(egocentric)의 RGB-D 영상, 에이전트의 위치 P_t, 그리고 목표 물체의 속성을 설명하는 자연어 문장 $sentence(g)$으로 구성된다. 에이전트가 선택 가능한 행동 $a \in A$은 A = {MoveForward, TurnLeft, TurnRight, LookUp, LookDown, Stop}의 6가지 이산화된 행동 집합 중 하나이다. 이때, 전진은 0.25 m, 각 회전은 60°로 고정된다. 에이전트가 시야 내에 목표 물체가 존재할 때 Stop 행동을 수행하면 작업 성공으로 간주하며, 목표 물체와의 거리가 d_s 이상인 경우, 잘못된 물체를 찾는 경우, 정해진 최대 시간 단계(max time-step)를 초과한 경우 등은 모두 작업 실패로 간주한다.

[Table 1] Goal object categories based on description types

Description Type	Goal Objects
(1) Appearance, (2) Spatial, (3) Hidden	AlarmClock, Apple, BaseballBat, BasketBall, Bowl, GarbageCan, HousePlant, Laptop, Mug, SprayBottle, Television, Vase
(4) Uncommon	GingerbreadHouse, EspressoMachine, Crate, ElectricGuitar, RiceCooker, LlamaWickerBasket, Whiteboard, Surfboard, Tricycle, GraphicsCard, Mate, ToyAirplane

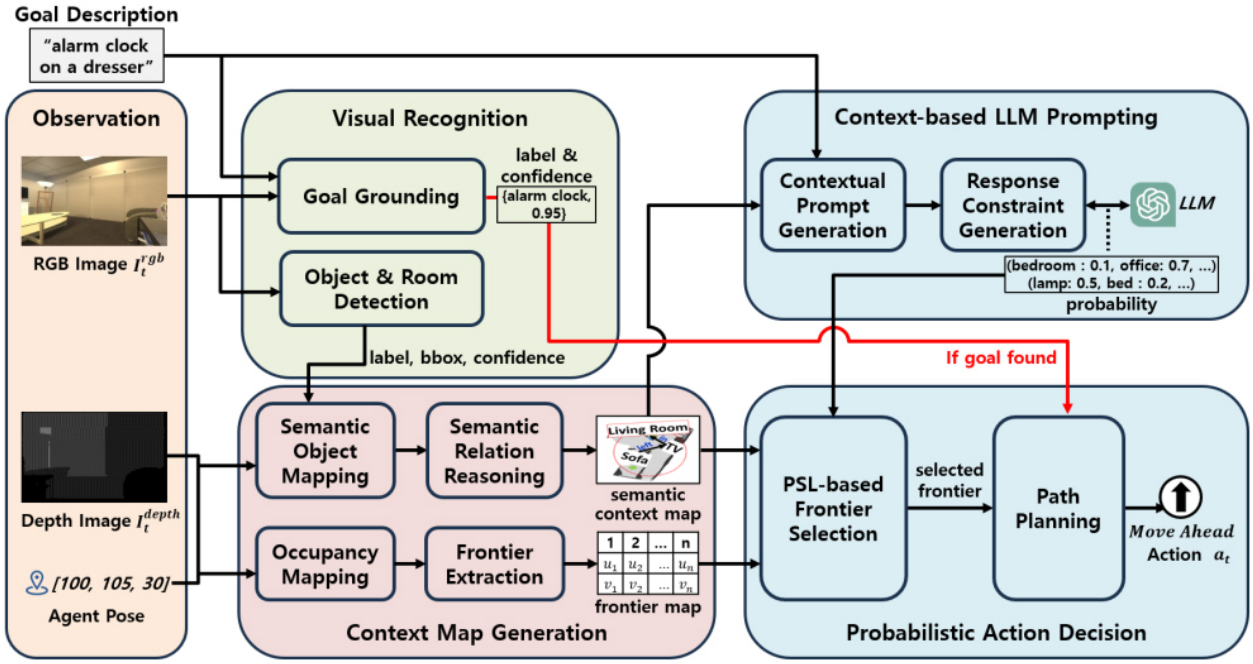
자연어 문장 $sentence(g)$는 목표 물체 g의 시각적 속성, 공간적 관계 등을 묘사한다. 이러한 문장은 다음 4가지 유형으로 나뉜다. (1) 외관 설명(Appearance description)은 '{size}, {color}, {material} {object}' 형태로 목표 물체의 시각적 속성에 대해 묘사한다. (2) 공간 설명(Spatial description)은 '{object} on top of {x}, near {y}, {z}, ...'의 형태로 목표 물체 주변의 다른 물체들과의 공간적 관계를 설명한다. (3) 숨겨진 물체 설명(Hidden description)은 "{object} under/in {x}"의 형태로 가구 아래 또는 안에 숨겨져 보이지 않는 목표 물체에 대해 묘사한다. (4) 흔하지 않은 물체 설명(Uncommon description)은 기존 벤치마크 데이터 집합에 등장하지 않는 "Cooker", "Espresso machine" 등 12개의 새로운 목표 물체 범주를 제공한다. 각 유형에서 주어지는 12개의 새로운 목표 물체 범주는 [Table 1]과 같다.

또한, (1) 외관 설명, (2) 공간 설명, (3) 숨겨진 물체 설명 유형에서는 목표 물체 범주마다 단 하나의 개체(instance)만 존재하는 기본(normal) 환경과, 동일 범주지만 시각적으로 구분되는 방해 개체(hindrance instance)가 추가된 방해물(distractor) 환경의 2종류로 나뉜다.

3.2 에이전트 모델의 구성

본 논문에서는 시각-언어 모델(VLM), 대규모 언어 모델(LLM)과 같은 사전 학습된 인공지능 기저 모델(AI foundation model)들을 활용하여 L-ZSON 작업을 수행하는 새로운 에이전트 모델 AML (Agent Model for L-ZSON)을 제안한다. 제안 모델 AML은 [Fig. 2]와 같이, (1) 영상 인식 모듈(Visual Recognition, VR) (2) 맥락 지도 생성 모듈(Context Map Generation, CMG), (3) 맥락 기반 대규모 언어 모델 프롬프팅 모듈(Context-based LLM Prompting, CLP), (4) 확률적 행동 결정 모듈(Probabilistic Action Decision, PAD)으로 구성된다.

영상 인식(VR) 모듈은 자유 형식의 언어 묘사 $sentence(g)$로 주어진 목표 물체 g를 실시간 RGB 영상 I_t^{rgb}에서 식별해내는 목표 물체 그라운딩(goal grounding) 과정과 물체와 방을 나



[Fig. 2] Architecture of the proposed agent model

열한 프롬프트(prompt)로부터 에서 환경을 구성하고 있는 방, 가구, 기타 물체들을 탐지해내는 물체와 방 탐지(object & room detection) 과정으로 구성된다.

맥락 지도 생성(CMG) 모듈은 에이전트의 이동 가능 영역 정보를 제공하는 점유 지도와 이로부터 탐험해야 할 영역 정보를 제공하는 프론티어 지도(frontier map)를 생성하는 과정, 그리고 환경 구성 요소 간의 공간적 배치 관계를 표현하는 의미적 맥락 지도를 생성하는 과정으로 구성된다.

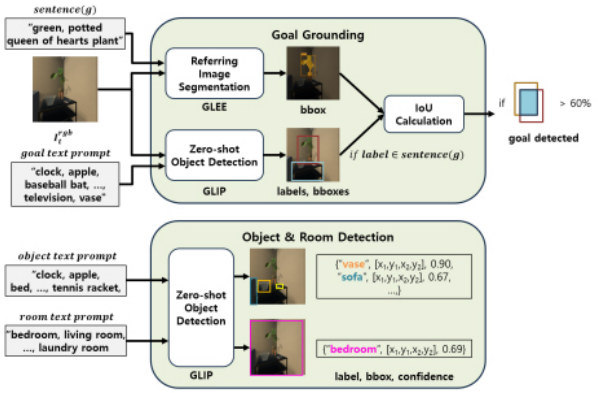
I_t^{rgb}맥락 기반 대규모 언어 모델 프롬프팅 모듈(CLP)에서는 CMG 모듈을 통해 생성된 의미적 맥락 지도로부터 구한 공간적 맥락정보를 활용하여 LLM에 질의할 프롬프트를 생성하고, 답변으로 물체들과 방들에 대한 목표 물체의 존재 가능성 점수를 받아내는 과정으로 구성된다.

프론티어 기반 탐험(frontier-based exploration) 알고리즘을 기반으로 하는 확률적 행동 결정 모듈(PAD)에서는 매순간 행동을 결정하기 위해 CMG 모듈의 프론티어 지도 및 의미적 맥락 지도, 그리고 CLP 모듈에서 구해낸 <목표-물체>, <목표-방> 간의 연관성 점수를 이용하여 최적의 프론티어 셀을 하나 선택하는 프론티어 선정(PSL-based frontier selection) 과정과 해당 프론티어 셀까지 도달하기 위한 최적의 행동 a_t을 연속적으로 선택하는 최적 경로 계획(path planning) 과정으로 구성된다.

탐색 이동 작업 시작과 동시에 에이전트는 매 순간 VR 모듈의 목표 물체 그라운딩에서 목표 물체 g의 관측 여부를 판단한다. 만약 목표 물체를 관측했다고 판단하면 위의 모든 과정을 생략하고 바로 PAD 모듈의 최적 경로 계획에서 목표까지 최단 거리로 이동하기 위한 단위 행동들을 수행한다. 반면 목표 물체를 아직 관측하지 못한 경우엔, 위의 모듈들을 순서대로 실행하며 탐험을 위한 행동들을 수행한다.

3.3 참조 영상 분할 모델을 이용한 목표 물체 그라운딩

제안 모델의 영상 인식 모듈(Visual Recognition, VR)에서는 자유 형식의 언어 묘사 $sentence(g)$로 주어진 목표 물체 g를 t시간의 RGB 영상 I_t^{rgb}에서 현재 목표 물체의 관측 여부를 판단하는 목표 물체 그라운딩(goal grounding) 과정과 I_t^{rgb}와 함께 물체와 방을 나열한 프롬프트로부터 시야 내의 물체들 $L_t = \{obj_1, ...obj_K\}$과 방 R_t을 탐지해내는 물체와 방 탐지(object & room detection) 과정으로 구성된다. [Fig. 3]은 해당 과정의 세부 단계들을 나타낸다. 먼저, 제안 모델의 핵심 요소인 목표 물체 그라운딩 과정에서는 자연어 문장으로 표현된 목표 물체 묘사를 이해하고 해당 물체를 입력 영상에서 식별해내는 구절 그라운딩(phrase grounding) 과정이 필요하다. 이와 관련하여 기존 연구[2,3]들에서 사용한 시각 인식 기저모델인 GLIP[5]은 추가 훈련 없이도 관측 영상 내 목표 물체를 효과적으로 탐지해내는 등 뛰어난 일반화 능력을 보인다. 하지만 본래는 비교적 길이가 짧은 탐지 대상 물체의 레이블(label)을 입력받도록 설계된 GLIP 텍스트 인코더(text encoder)의 제한된 입력 토큰 길이(input token size)로 인해, “red apple on the wooden table”과



[Fig. 3] Visual recognition

같이 목표 물체 자연어 묘사가 길어지면 목표 물체를 나타내는 apple과 같은 핵심 단어와 table과 같은 부속 단어들을 구분해 내기 어려워, 간혹 핵심 단어 apple이 가리키는 사과가 아니라 부속 단어 table이 가리키는 탁자를 탐지해내는 문제가 존재한다. 반면에, 대표적인 참조 영상 분할(referring image segmentation) 모델인 GLEE는 영상 내의 특정 대상을 지칭하기 위해 사용되는 자연어 묘사인 참조 표현(referring expression)을 이해하고, 이 참조 표현에 가장 부합되는 물체 영역을 입력 영상에서 정확하게 분할해내는 기능을 제공한다. 하지만 GLEE 모델은 참조 영상 분할 모델의 특성상 참조 대상인 목표 물체가 영상 내에서 관측되지 않을 때에도 영상 안의 다른 물체를 목표 물체로 식별해내는 약점을 가지고 있다. 이러한 특성은 작업 대부분의 시간 동안 목표 물체가 시야 내에 들어오지 않는 L-ZSON 작업에서는 치명적인 문제가 될 수 있다. 이를 보완하기 위해 제안 모델의 목표 물체 그라운딩 모듈에서는 GLEE와 GLIP을 상호 보완적으로 함께 이용하여 입력 영상에서 목표 물체를 식별해내는 신뢰도를 개선하고자 한다.

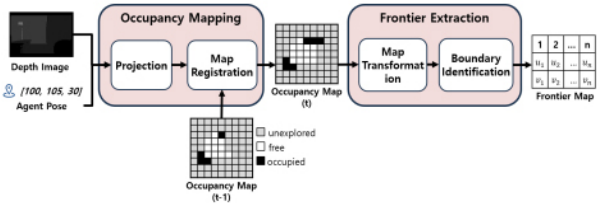
먼저 GLEE 모델에 RGB 입력 영상 I_t^{rgb}와 'green, potted queen of hearts plant'와 같은 현재 목표의 속성을 묘사한 문장 $sentence(g)$을 입력하여 분할 영역에 대한 경계상자(bounding box)를 구해낸다. 이와 동시에, GLIP 모델로부터 I_t^{rgb}와 12개의 개별 목표 물체들을 "."으로 구분한 텍스트 프롬프트(text prompt)를 입력하여 레이블(label)과 경계상자를 구해낸다. GLIP의 예측 레이블들과 문장 속 목표 단어 g와 비교하여 목표 물체에 해당하는 경계상자를 선정하고, 두 모델의 경계상자 간의 IoU (Intersection of Union)가 60% 이상으로 계산되면 목표 물체를 발견한 것으로 판단한다. 목표 물체를 발견하면, GLEE 모델의 경계상자 중심점(center point)에 해당하는 픽셀의 깊이값으로부터 목표 물체의 전역적 위치를 계산해내고, 최단 거리 알고리즘(A*)을 통해 해당 위치로 이동한다.

한편 물체와 방 탐지 과정에서는 GLIP 모델만을 이용한다. 앞서 언급한 GLIP의 일반화 능력은 에이전트가 관측한 시야 내의 물체뿐만 아니라 방과 같은 공간적 정보를 추론하는 데도 효과적으로 적용될 수 있다. 이는 에이전트가 작업 환경을 이해하고 해석하는 데 매우 중요한 역할을 수행한다. 해당 과정에서는 먼저 GLIP모델에 RGB 입력 영상 I_t^{rgb}와 함께 목표 물체에 해당하지 않는 환경 물체 30개를 '.'의 구분자와 함께 나열한 물체 텍스트 프롬프트를 입력하여 영상 내 물체들에 대한 레이블과 경계상자, 그리고 신뢰도(confidence)를 얻어낸다. 9종류의 방 탐지도 물체와 동일하게 진행한다. 이후 물체와 방 모두 GLIP의 탐지 신뢰도가 0.61 이상이라면, 출력된 레이블의 물체들을 관측했다고 판단한다.

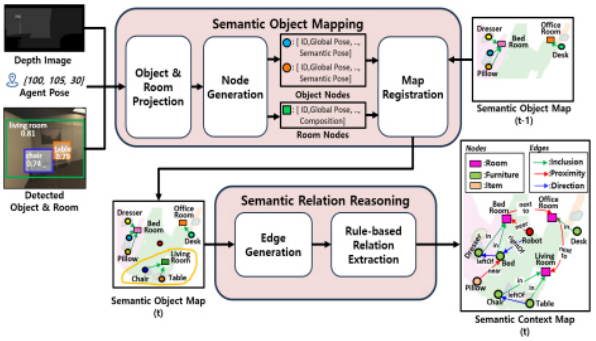
3.4 공간적 맥락 지도 생성

제안 모델의 맥락 지도 생성(CMG) 모듈은 에이전트의 이동 가능 영역 정보를 제공하는 점유 지도와 이로부터 탐험해야 할 영역 정보를 제공하는 프론티어 지도(frontier map)를 생성하는 과정, 그리고 환경 구성 요소 간의 공간적 배치 관계를 표현하는 의미적 맥락 지도를 생성하는 과정으로 구성된다. 첫 번째로 점유 지도와 프론티어 지도 작성 과정에서는 [Fig. 4]와 같이, 깊이 영상 I_t^{depth}와 에이전트의 현재 위치 P_t를 입력받아 각 셀의 장애물 점유 여부를 표현하는 점유 지도를 생성하고, 미탐험 영역을 식별하기 위해 탐험 영역과 미탐험 영역 간 경계에 해당하는 셀들을 추출한 프론티어 지도를 생성한다.

두 번째 단계는 [Fig. 5]와 같이, VR 모듈에서 탐지된 환경 구성 요소들의 범주 정보와 전역적 위치 정보를 활용하여 노드만으로 구성되는 그래프 형태의 의미적 물체 지도(semantic object map)를 작성한다. 이때 각 노드는 작업 환경을 구성하는 물체와 방으로 구성되며, 각 노드는 전역적 위치(global location), 물체의 방 소속을 나타내는 의미적 위치(semantic location), 범주(category) 및 탐지 신뢰도(detection confidence)와 같은 주요 속성을 갖는다. 이후, 의미적 관계 추론(semantic relation reasoning) 과정을 통해 의미적 물체 지도의 노드 간의 전역적인 공간적 관계를 표현한 의미적 맥락 지도(semantic context map)를 작성한다. 이때 각 노드는 물체와 방 노드뿐만 아니라 에이전트 노드까지 추가한다. 각 노드를 연결하는 간선은 물체, 방, 에이전트



[Fig. 4] Occupancy and frontier map generation



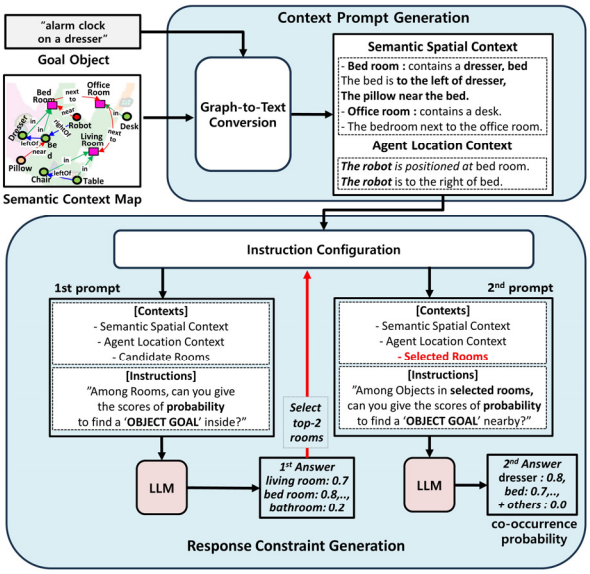
[Fig. 5] Semantic context map generation

간의 상호작용과 공간적 관계를 표현한다. 이와 같은 추론 과정을 통한 제안 모델의 의미적 맥락 지도는 물체 또는 방 종류만 표현되거나, 물체 간의 위치 관계로 표현이 제한되는 기존 연구들[3,15,16]과 달리, 노드와 관계 간선들을 유형별로 나누고 계층적으로 환경을 이해함으로써, 에이전트가 복잡한 공간적 관계를 명확히 파악하고 효과적인 이동 계획 수립에 도움을 줄 수 있다.

노드와 관계 간선은 각각 4가지 유형으로 구분한다. 먼저, 3가지 노드는 각각 다음과 같다. 작업 환경 내에서 Room 노드는 “bedroom”, “living room”과 같은 공간 단위를 표현하고, Furniture 노드는 “sofa”, “table”과 같이 크기가 크고 고정된 가구 물체를 표현한다. 반면에 Item 노드는 “pillow”, “lamp”과 같이 상대적으로 작고 이동 가능한 기타 물체들을 표현한다. Agent 노드는 이동 작업을 수행하는 로봇 에이전트의 위치를 표현한다. 이러한 크기 및 역할에 따른 계층적 구분은 작업 환경 내의 다양한 구성 요소들과 그들 사이의 복잡한 공간적 관계를 정확히 파악하는 데 도움을 줄 수 있다. 한편, 간선의 경우엔 서로 다른 3가지 공간적 관계, 즉 포함 관계(inclusion relation), 방향 관계(direction relation), 근접 관계(proximity relation)들로 구분한다. 포함 관계는 <dresser, in, bedroom>과 같이, Room 노드와 해당 영역에 위치하는 Furniture 노드 또는 Agent 노드와의 관계를 의미하며 이는 실내 환경의 구조와 각 방의 역할을 이해하도록 돕는 역할을 한다. 방향 관계는 <table, left_of, chair>, <agent, right_of, bed>와 같이, Furniture 노드들 또는 그들과 Agent 노드 간의 상대적 위치와 방향을 나타내며, 이는 물체들의 실내 배치와 상호작용을 명확히 파악하는 데 도움이 된다. 마지막으로 근접 관계는 <pillow, near, bed>, <bedroom, next_to, living room>와 같이, 4종류의 모든 노드들간의 위치에 기반한 인접성을 표현한다.

3.5 대규모 언어 모델 프롬프팅

제안 모델의 맥락 기반 대규모 언어 모델 프롬프팅 모듈



[Fig. 6] Context-based LLM prompting

(CLP)에서는 [Fig. 6]과 같이, CMG 모듈의 의미적 맥락 지도로부터 구한 공간적 맥락정보를 활용한다. 이때, 현재 에이전트 위치를 기준으로 주변 환경을 구성하는 물체들과의 공간적 관계를 설명한 에이전트 위치 맥락정보(agent location context), 환경 물체들 간의 전역적 공간 관계와 방과 방 사이의 연결 관계를 설명한 의미적 공간 맥락정보(semantic spatial context)를 기반으로 두 가지 프롬프팅이 이루어진다.

첫 번째 프롬프팅에서는 [Fig. 6]의 1st prompt와 같이, 에이전트 위치 맥락정보와 의미적 공간 맥락정보를 바탕으로 전체 후보 방들에서 목표 물체를 찾을 확률값을 응답받는다. 이러한 프롬프팅 전략은 현재 관측 물체 또는 방과 물체 후보 등 제한된 맥락정보만을 LLM에 전달하는 기존 모델[1-3]들과 달리, 에이전트의 주변 환경에 대한 공간적 맥락정보뿐만 아니라 과거 탐험 이력으로부터 얻은 맥락정보도 함께 고려함으로써, 에이전트가 탐색해야 할 프론티어 영역을 우선순위에 따라 결정하는 데 매우 유용하다. 두 번째 프롬프팅에서는 [Fig. 6]의 2nd prompt와 같이, 첫 번째 응답으로부터 받은 상위 두 개의 방 범주 정보를 추가 전달하여 LLM으로부터 선정된 방들에 포함된 환경 물체들과 목표 물체가 동시 출현할 확률값을 응답받는다. 만약 선정된 방이 에이전트가 이전에 탐험하지 않은 방일 경우 해당 방이 미탐험 영역임을 명시한다. 목표 물체와의 연관성 점수를 추천받기 위해 두 번의 프롬프팅을 독립적으로 수행하는 ESC[3] 모델과 달리, 제안 모델에서는 제약 조건을 추가한다. 즉, 에이전트가 실제 경험한 환경 구성을 기반으로 추천 가능한 물체들의 범위를 축소함으로써 일반 상식에 대한 의존도를 낮추고, 탐험 과정에서의 불확실성을 줄이는 데 도움을 줄 수 있다.

3.6 확률적 행동 결정

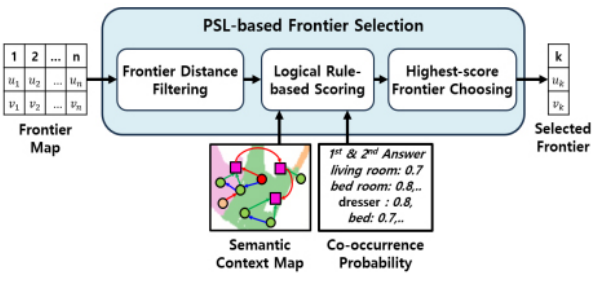
확률적 행동 결정(PAD) 모듈에서는 프론티어 기반 탐험(FBE) 알고리즘을 사용한다. 그러나 기존의 거리 기반 프론티어 탐색 방식은 작업 환경의 맥락정보를 충분히 반영하지 못하여 상식에 어긋날 수 있으며, 비효율적인 경로를 통해 이동하게 되므로 제안 모델은 상식을 기반으로 프론티어를 선택하는 확률적 소프트 논리(Probabilistic Soft Logic, PSL) 방식[3]을 도입한다. PSL은 프론티어 선택 시 에이전트로부터의 거리 d_i뿐만 아니라, 프론티어 주변의 물체들과 공간 정보를 모두 고려하여 목표 물체가 등장할 가능성이 높은 물체 또는 방이 존재하는 프론티어를 선택한다. 하지만 이를 만족하는 프론티어는 여러 개가 해당될 수 있으므로 확률적으로 규칙을 정의한다. 규칙은 특정 도메인 내 개체 간의 관계와 속성을 원자(atoms)로 정의하고 이를 가중치가 부여된 일차 논리절과 선형 산술 부등식으로 표현한다. [Fig. 7]은 해당 과정의 세부 단계들을 나타낸다.

먼저, 프론티어 거리 필터링(frontier distance filtering) 과정에서는 앞서 점유 지도로부터 추출한 프론티어 셀들의 전역적 셀 위치(index)를 입력으로 받아, 현재 에이전트 위치로부터 특정 거리($1.2m$) 이내인 프론티어들을 제외한다. 이는 에이전트가 넓은 영역을 효과적으로 탐험할 수 있도록 하기 위함이다. 논리 규칙 기반 점수 계산(logical rule-based scoring) 과정에서는 의미적 맥락 지로로부터 관측 물체/방의 위치와 탐지 신뢰도, 거리에 따라 필터링된 프론티어 셀 위치(index) 리스트, 그리고 LLM으로부터 받은 <목표-물체>, <목표-방> 간의 연관성 점수를 입력으로 받아 각 프론티어마다의 목표 물체가 존재할 가능성에 대한 점수를 계산해낸다. [Fig. 8]은 해당 과정의 세부 단계들을 나타낸다. 해당 과정에서 PSL을 사용하며, 제안 모델에서 구현된 네 가지 주요 규칙은 다음과 같다. 참고로, 파라미터 w는 규칙의 가중치로, 모델에서 상대적 중요성을 정량화하는 값이다.

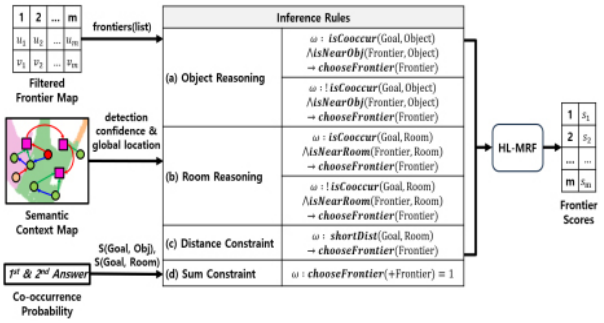
[Fig. 8]의 (a) 물체 추론(object reasoning)은 에이전트가 목표 물체 주변에 나타날 가능성이 있는 일부 환경 물체 근처의 프론티어를 탐색하도록 유도하기 위해 다음과 같은 2가지 규칙을 사용한다. 해당 규칙에는 세 개의 원자(atoms)가 사용되는데, LLM에서 예측한 (Goal, Object) 쌍에 대한 동시 출현(co-occurrence) 점수 $S(G|o_i)$값인 $isCooccur(Goal, Object)$, 해당 프론티어에서 일정 거리 이내에 있는 탐지 물체들의 GLIP 신뢰도 값인 $isNearObj(Frontier,\ Object)$, $chooseFrontier(Frontier)$ 등 3가지 요소를 사용한다. $isNearObj(Frontier, Object)$에선 거리 임계치 이상인 경우 또는 탐지되지 않은 물체들은 모두 0값을 갖는다. 이뿐만 아니라 에이전트가 목표 물체 주변에 있을 가능성이 낮은 일부 환경 물체 근처 프론티어로의 이동을 방지하기 위해 $!isCooccur(Goal, Object)$를 이용한 부정 규칙도 함께 사용한다.

[Fig. 8]의 (b) 방 추론(room reasoning)은 물체 추론과 유사하게 두 가지 규칙을 정의한다. 즉, 에이전트가 목표 물체가 나타날 가능성이 있는 방 안이나 방 근처의 경계를 탐색하도록 유도한다. 또, 목표 물체가 나타날 가능성이 낮은 방 안 또는 근처는 탐험하지 않도록 제한한다. [Fig. 8]의 (c) 거리 제약(distance constraint)은 에이전트가 가까운 프론티어를 탐험하도록 장려하기 위한 최단 거리 규칙이다. 이는 에이전트와 거리가 가장 짧은 프론티어를 선택하여 에이전트가 근처 영역에서 더이상 탐험할 것이 없을 때까지 한 영역을 계속 탐험하도록 유도하기 위함이다. [Fig. 8]의 (d) 합산 제약(sum constraint)은 모든 프론티어를 선택한 점수의 합을 1로 제한하기 위한 확률 제약 조건으로, 관측되지 않은 모든 변수가 1이 되는 경우들을 방지하고 모든 프론티어들이 선택될 가능성을 부여한다. 기본적으로 PSL 모델의 모든 규칙에 동일한 가중치를 사용한다.

이후 4가지 규칙에 의한 chooseFrontier (Frontier) 값들은 앞서 언급한 힌지 손실 마르코프 랜덤 필드(HL-MRF) 방식에 의해 최종 점수가 매겨진다. 해당 과정은[3] 연구의 방식을 따른다. 마지막으로 가장 높은 점수를 갖는 프론티어 하나를 선택하게 되면, 목표 프론티어 셀에 도착할 때까지 매순간 에이전트의 위치에서부터 해당 셀까지의 A* 알고리즘을 기반으로 최단 거리를 계산하고, 가장 최적의 행동 a_t를 선택한다.



[Fig. 7] PSL-based frontier selection



[Fig. 8] Logical rule-based scoring

4. 구현 및 실험

4.1 모델 구현

본 논문의 제안 모델은 Python 딥러닝 라이브러리인 Pytorch를 이용해 구현하였으며, 모든 성능 평가 실험들은 Intel Xeon Silver 4210R CPU와 GeForce RTX 3080 Ti GPU 2개가 장착된 컴퓨터 환경에서 동일하게 수행되었다. 그리고 제안 모델의 성능 평가에는 3차원 실사(photo-realistic)의 대규모 실내 시뮬레이션 환경을 제공하는 RoboTHOR[10] 기반의 벤치마크 데이터 집합인 PASTURE[1]를 이용하였다. RoboTHOR에서 제공하는 15개의 검증(validation)용 환경에서 에피소드 1,800개로 측정한 작업 성능을 비교하였다. 에피소드 성공 판단을 위한 에이전트와 목표 물체 간의 거리 임계값(threshold)은 1.5 m, 최대 시간 단계(max time-step)는 250으로 설정하였다. 이때, 외관 설명, 공간 설명, 흔하지 않은 물체 설명 유형에서는 목표 물체를 기준으로 관측 여부 및 에이전트와의 거리를 고려하여 탐색 이동 작업의 성공을 판단하며, 숨겨진 물체 설명 유형에서는 목표 물체를 숨긴 물체를 기준으로 성공 여부를 판단한다.

제안 모델에서 사용한 GLIP 및 GLEE, 그리고 LLaMA2는 모두 제로-샷 작업을 위하여 RoboTHOR 환경에서 별도의 훈련을 진행하지 않았다.

4.2 성능 평가 실험

본 논문에서는 제안 모델의 우수성을 입증하기 위해, (1) 목표 물체 그라운딩 방식에 따른 분할 성능 비교, (2) 전역적 맥락 지도 구성 방식에 따른 작업 성능 비교, (3) LLM 프롬프팅 방식에 따른 작업 성능 비교, (4) 다른 최신 모델들과의 비교 등 총 4가지 정량 평가 실험들을 진행하였다. 이와 더불어 (5) 목표 물체 그라운딩 방식에 따른 정성적 비교와 (6) 최신 모델들 대비 제안 모델의 장점과 한계점을 분석하기 위한 비교 등 2가지 정성 평가 실험도 함께 진행하였다. 특히, 목표 물체 그라운딩 모듈의 성능 평가 실험에서는 목표 탐지 성능을 집중적으로 평가하기 위하여 따로 수집한 RGB 영상과 이에 해당하는 경계상자(bounding box) 쌍 또는 이진 분할 마스크(binary segmentation mask) 쌍을 사용하였다. 이때, 사전 정의된 PASTURE 에피소드마다 FBE 알고리즘을 이용하여 각 목표 g에 대한 RGB 영상, 정답 경계상자, 그리고 정답 분할 영상을 수집하였다. 각 방식은 자연어 문장 $sentence(g)$이 주어졌을 때, 영상 내에서 해당 물체를 올바르게 그라운딩할 수 있는지 측정하기 위해 성능 평가 지표(metric)는 oIoU (overall Intersection over Union, %), Acc (Accuracy, %)를 이용하였다. oIoU는 참조 영상 분할 작업에서 주로 채택하는 평가 지표로, 식 (1)과 같이 계산한다. 이때 n은 각 데이터 집합마다의 평가 영상 개수, U와 I는 각각 정답 마스크와 예측 마스크 간의 합집합과 교집합을 의미한다. 즉, 모든 평가 영상에 대해 교집합의 면적과 합집합의 총면적으로 나눈 누적값을 의미한다. Acc는 레이블 예측에 대한 이진 크로스 엔트로피(binary cross-entropy)로, Hidden 데이터 집합에서 RGB 영상 내 목표 물체가 존재하지 않음을 판단하는 능력을 평가하기 위해 사용하였다.

$$oIoU = \sum_{i=1}^{n} \frac{I(i)}{U(i)} \tag{1}$$

이동 탐색 작업 성능 평가를 위해서는 SR (Success Rate, %)과 SPL (Success rate weighted by Path Length, %) 2가지를 이용하였다. 단순 성공률을 나타내는 SR에 비해, SPL은 작업 성공률뿐만 아니라 이동 경로 길이도 함께 고려한다는 차이점이 있다. SR은 식 (2)와 같이 계산하며, SPL은 식 (3)과 같이 계산한다. 식 (2)의 N은 전체 에피소드의 개수이며, S_i는 i번째 에피소드에서의 성공(success) 여부 (1 혹은 0)를 나타낸다. 식 (3)의 l_i는 목표 물체까지의 최단 이동 경로, p_i는 에이전트의 실제 이동 경로 길이이다. 두 가지 지표는 모두 값이 클수록 긍정적인 성능을 의미한다.

$$SR = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{2}$$

$$SPL = \frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)} \tag{3}$$

첫 번째 실험은 참조 표현 이해/참조 영상 분할 작업에서 제안 모델에서 채용한 GLEE 모델과 GLIP 모델을 함께 이용한 목표 물체 그라운딩 방식의 긍정적인 효과를 입증하기 위한 실험이다. 이 실험에서는 목표 물체 그라운딩 모듈만 독립적으로 평가하였으며, 제로-샷 개방형 어휘 물체 탐지기인 (a) OWL-ViT[1], (b) GLIP[2,3], 그리고 참조 영상 분할 모델인 (c) GLEE[9], 그리고 제안 모델과 같이 (d) GLEE+GLIP을 사용한 경우로 구분하여 각각의 성능을 비교하였다.

[Table 2]의 실험 결과를 살펴보면, 모든 데이터 집합에서 본 논문에서 제안한 GLEE+GLIP의 방식을 채용한 경우에 가장 높은 oIoU, Acc 성능을 보이는 것을 확인할 수 있다. 구체적으로 살펴보면, Uncommon 데이터 집합의 경우, 제안 방식인 GLEE+GLIP이 제로-샷 물체 탐지기인 OWL-ViT, GLIP, 그리고 참조 영상 분할 모델인 GLEE보다 각각 10.84배, 3.69배, 1.13배 높은 IoU 성능을 보인다. 일반적이지 않은 물체 레이블

[Table 2] Performance comparison with different goal grounding methods

Goal Grounding Method		Pre-training Dataset	oIoU(%) on PASTURE					Acc(%)	
			Uncommon	Spatial		Appearance		Hidden	
			normal	normal	distract	normal	distract	normal	distract
Zero-shot OD	OWL-ViT[1]	COCO, Open-Image	3.43	3.42	2.71	5.74	3.11	38.23	23.67
	GLIP[2,3]	FourODS, GoldG, Cap4M	10.06	9.66	3.71	16.11	23.38	40.39	38.86
RIS Model	GLEE[9]	Objects365, OpenImages, COCO, LVIS, Visual Genome	33.03	48.74	42.48	53.62	42.09	22.88	21.76
-	GLEE + GLIP	-	**37.17**	**53.26**	**46.36**	**58.54**	**56.81**	**99.95**	**99.41**

임에도 올바르게 탐지해내는 것을 통해 제안 방식에서 추가 채용한 GLEE 모델의 높은 범용성을 확인할 수 있다. 추가로, 영상 단위(image-level)로 처리하는 물체 탐지기인 OWL-ViT보다 물체 단위(object-level) 탐지기인 GLIP이 oIoU가 2.93배 더 높은 것을 확인할 수 있다. 이를 통해 GLIP이 텍스트를 이미지 내 구체적인 요소와 직접 연결하여 정확한 목표 위치를 식별해내는 데 효과적임을 확인할 수 있다. 한편, Appearance-normal, Spatial-normal 데이터 집합의 결과를 살펴보면, OWL-ViT와 GLIP 대비 GLEE+GLIP 방식이 oIoU가 각각 15.57배, 5.31배, 1.09배 높은 성능을 보이는 것을 확인할 수 있다. 특히, 목표 물체를 표현하는 명사와 함께 다양한 물체들이 함께 등장하는 Spatial에서 높은 성능 향상률을 보이는데, 이는 제로-샷 물체 탐지기들은 주어에 해당하는 물체만을 올바르게 판단해내지 못하여 목표 물체를 관측했음에도 탐지해내지 못하거나, 문장 내 함께 등장하는 환경 물체를 목표 물체로 오인하는 경우가 다수 존재하기 때문으로 보인다. 마지막으로 Hidden 데이터 집합의 경우, 목표 물체가 존재하지 않음을 판단하기 위한 데이터 집합으로, 제안 방식이 normal과 distractor 모두 매우 높은 Acc 성능을 보임을 확인할 수 있다. 특히, GLEE의 경우에 가장 낮은 Acc 성능을 보이며, 이는 제안 방식이 GLIP을 이용하여 필터링하기 때문에 관측 영상 내 목표 물체가 없음을 판단하는 것과 달리 목표 물체가 존재하지 않음에도 반드시 하나의 물체를 잡아내기 때문에 목표 물체가 아님에도 목표라고 탐지해내는 거짓 양성(false positive)에 해당하는 오탐지가 많기 때문이라고 볼 수 있다.

두 번째 실험에서는 환경 구성 요소들 간의 공간적 배치 관계를 추론하는 제안 모델의 의미적 맥락 지도 생성 방식의 우수성을 입증하기 위한 실험이다. 이 실험에서는 에이전트의 현재 위치로부터 점유 여부를 나타내는 (a) 점유 지도(occupancy grid map)만을 이용하는 경우, 관측된 환경 구성 요소들의 범주 정보만을 포함한 (b) 의미적 물체 지도(semantic object context map)를 이용하는 경우, 그리고 구성 요소 간의 공간적 위치 관계들을 추가로 활용한 (c) 의미적 맥락 지도(semantic context map)를 이용하는 경우를 서로 비교하였다. 점유 지도를 이용한 (a)의 경우에는 최단 거리 기반의 프론티어 탐색 전략만을 적용하였고, 의미적 지도를 이용한 (b), (c)의 경우에는 환경 구성 요소에 대한 의미 정보를 바탕으로 LLM 프롬프팅 과정과 확률적 행동 결정 방법을 적용하였다.

[Table 3]의 실험 결과를 살펴보면, 대부분의 실험 결과에서 제안 방식인 (c)가 가장 높은 SR, SPL 성능을 보이는 것을 확인할 수 있다. 구체적으로 비교해보면, 점유 지도만을 이용한 (a) 경우보다 환경 내 물체와 요소들의 종류 및 탐지 신뢰도 등의 의미적 정보를 포함한 지도를 이용한 (b), (c)의 경우들이 더 높은 SPL 성능을 보이는 것을 통해 효율적인 경로로 목표 물체까지 이동함을 확인할 수 있다. 또, 환경 구성 요소들의 범주 정보만을 이용한 (b)와 그들 간의 공간 관계 정보를 추가로 이용한

[Table 3] Performance comparison with different spatial context maps

Spatial Context Map	SR(%) / SPL(%) on PASTURE													
	Uncommon		Spatial				Appearance				Hidden			
	normal		normal		distract		normal		distract		normal		distract	
	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
occupancy map[1,2]	30.83	16.40	31.11	16.44	20.52	13.31	30.38	16.32	21.11	12.87	3.61	1.33	**3.05**	**1.43**
semantic object map[3]	32.50	23.50	30.27	16.86	20.27	12.67	31.66	20.27	20.55	12.03	**5.14**	**2.06**	2.50	1.08
semantic context map	**34.44**	**24.62**	**34.38**	**25.92**	**27.50**	**17.38**	**33.32**	**25.02**	**28.97**	**17.59**	1.83	0.99	2.22	0.98

(c)의 경우를 비교해보면, 특히 Spatial-normal에서는 1.14배, Spatial-distract에서는 1.27배의 높아진 SR 성능을 보이는데, 이는 물체들 간의 상대적 위치와 방향성, 그리고 그들이 서로 얼마나 가까이 있는지와 같은 공간적 관계 정보가 경로 계획과 탐색 효율성을 개선하는 데 도움을 줄 수 있다는 것을 확인할 수 있었다. 위의 결과들을 통해 제안 모델의 의미적 맥락 지도 생성 방식이 작업 성공과 효율성에 미치는 효과를 확인할 수 있었다.

세 번째 실험에서는 제안 모델의 맥락 기반 대규모 언어 모델 프롬프팅 모듈(CLP)에서 사용한 두 단계로 구성된 LLM 프롬프팅 방식의 우수성을 입증하기 위한 실험이다. 프롬프트 유형(prompt type)과 LLM 입력 맥락(input context)의 조합에 따른 4가지 방식을 각각 비교하였다. 프롬프트 유형에서 단일(single) 프롬프팅 방식은 방과 물체에 대한 프롬프트가 독립적으로 처리되는 반면, 연쇄(chain) 프롬프팅 방식은 첫 번째 프롬프트에서 특정 방에 대한 목표 물체 발견 확률을 추정한 뒤, 그 결과를 바탕으로 추가적인 질문을 구성하여 목표 물체와 연관성이 낮은 물체들을 배제하는 방식이다. 또한, LLM 입력 맥락은 LLM에게 환경 내 물체와 방에 대한 범주 레이블 정보만을 제공하는 물체와 방 레이블(object & room label)의 경우, 추가로 환경 구성 요소들 간의 공간적 관계(spatial relation)를 제공하는 경우를 나타낸다. 4가지 방식 모두 LLM의 답변으로는 각 상황에서 목표 물체의 존재 가능성 점수를 요청하였다. [Table 4]의 실험 결과를 살펴보면, 모든 데이터 집합에서 제안된 LLM 프롬프팅 방식을 적용한 경우에 가장 높은 SR 성능을 보이는 것을 확인할 수 있다. 구체적으로 살펴보면, Uncommon 데이터 집합의 경우, 제안 모델의 연쇄 프롬프팅 방식이 단일 프롬프팅 방식에 비해 물체와 방에 대한 범주 레이블 정보만 제공한 물체와 방 레이블 경우에서는 1.094배, 1.098배 높은 SR, SPL 성능을 보였고, 공간적 관계 경우에서는 1.14배, 1.09배 더 높은 성능을 보였다. 이로써, LLM의 첫 번째 응답을 기반으로 목표 물체와 연관성이 높은 환경 물체들을 선별하는 것이 탐험 범위를 축소하고, 탐험 과정에서의 불확실성을 줄이는데 효과적임을 확인할 수 있었다.

공간적 관계 경우를 물체와 방 경우들과 비교했을 때, Spatial-normal 데이터 집합에서 SR과 SPL 측면에서 연쇄 프롬프팅의 경우에서는 1.026배, 1.084배, 단일 프롬프팅에서는 1.03배, 1.14배의 더 높은 성능을 보이는 것을 확인할 수 있다. 이는 방에 속한 물체들의 종류뿐만 아니라, 방들 간의 연결 관계 및 물체들 간의 거리, 배치, 그리고 방향 관계를 LLM에게 알려줌으로써 작업 환경을 더 잘 이해하는 데 효과가 있음을 확인할 수 있었다. 또한, 같은 범주의 목표 물체와 시각적으로 구분되는 방해 요소(distractor)가 포함된 복잡한 환경인 Spatial-distract 데이터 집합에서도 가장 높은 SR과 SPL 성능을 보였다. 이는 “apple near a laptop” 시나리오에서 에이전트가 “apple” 물체뿐만 아니라 주변 물체인 “laptop”이 일반적으로 어느 방에 위치하는지 파악하고, 이를 통해 잠재적으로 혼선을 줄 수 있는 “apple on a coffee table”을 구분하여 효율적인 경로로 목표 물체까지 이동하는 제안 방식의 우수성을 확인할 수 있었다. 반면, Appearance-normal 데이터 집합의 결과를 살펴보면, 연쇄 프롬프팅 방식에서는 공간적 관계 경우가 1.04배, 1.23배 향상된 SR과 SPL 성능을 보이지만, 단일 프롬프팅 방식에서는 SR에서는 1.17배, SPL에서는 1.19배 더 낮은 성능을 보였다. 이는 단일 프롬프팅 시, 에이전트가 탐험 과정에서 탐지한 모든 방과 물체의 공간적 맥락을 LLM에 제공하기 때문에, 과도한 정보량으로 인해 오히려 추론에 방해가 되어 SR, SPL 성능이 떨어진 것으로 보인다. 하지만 제안 방식인 연쇄 프롬프팅 방식에서는 목표 물체와 같은 공간에 있을 법한 물체들을 미리 선별하기 때문에 정보 과부하를 줄일 수 있으므로 더 높은 성능을 보인다고 할 수 있다. 위의 결과들을 통해 제안 모델의 LLM 프롬프팅 방식이 작업 성공과 효율성에 미치는 효과를 확인할 수 있었다.

네 번째 실험은 최근에 발표된 다른 L-ZSON 모델(SOTA model)들과의 비교를 통해 제안 모델의 우수성을 입증하기 위한 실험이다. 해당 실험에서는 (a) 초기 L-ZSON 모델로써 목표 그라운딩 모델로 OWL-ViT를 활용한 CoW[1] (b) 목표 그라운딩 모델로 GLIP을 활용하였으며 YOLO를 기반으로 탐지된 환경 물체 레이블을 LLM에 제공하여 관련 물체로 향하는 이동

[Table 4] Performance comparison with different LLM prompting methods

LLM Prompting Method		SR(%) / SPL(%) on PASTURE													
		Uncommon		Spatial				Appearance				Hidden			
Prompt Type	LLM Input Context	normal		normal		distract		normal		distract		normal		distract	
		SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
single	object & room label	29.70	21.41	30.12	15.18	18.00	11.98	27.94	16.90	23.33	14.58	1.17	0.78	1.48	0.44
	spatial relation	30.12	22.39	30.94	17.33	19.42	12.25	27.22	14.16	24.79	15.69	1.26	0.86	2.17	0.93
chain	object & room label	32.50	23.50	30.32	15.16	20.27	12.67	31.66	20.27	23.55	15.03	1.45	0.93	1.96	0.79
	spatial relation	**34.44**	**24.62**	**31.11**	**16.44**	**27.50**	**17.38**	**33.32**	**25.02**	**28.97**	**17.59**	**1.83**	**0.99**	**2.22**	**0.98**

결정을 수행한 LGX[2] (c) 목표 물체와 환경을 구성하는 방, 기타 물체 탐지에 모두 GLIP을 사용하고, LLM으로부터 받은 목표 물체와 환경 물체, 그리고 방들 간의 동시 출현 확률값을 활용한 PSL을 기반으로 다음 탐험 지점을 결정하는 ESC[3] 등 총 3개의 최신 모델들과 (d) 제안 모델의 성능을 비교하였다.

[Table 5]의 실험 결과를 살펴보면, 본 논문의 제안 모델이 비교 대상인 다른 모든 모델들에 비해 대부분의 데이터 집합에서 가장 높은 성능을 보여주었다. 구체적으로 제안 모델은 CoW, LGX, ESC와 차례대로 작업 성공률을 나타내는 SR을 비교했을 때, Uncommon에서 1.097배, 2.04배, 1.069배, Spatial에서 1.99배, 1.42배, 1.13배, Appearance에서 1.19배, 1.38배, 1.08배의 성능 개선을 보여주었다. 작업 효율성을 나타내는 SPL 지표에서도 Uncommon에서 1.23배, 2.03배, 1.26배, Spatial에서 1.81배, 1.006배, 1.03배, Appearance에서 1.62배, 1.42배, 1.32배의 높은 성능 향상률을 보여주었다. 이와 같은 실험 결과를 바탕으로 본 논문에서 제안한 모델의 효과와 우수성을 확인할 수 있었다. 먼저, SPL 비교 시, LGX 낮은 SPL 성능을 보이는데, 이는 주로 관측한 물체를 기준으로 이동하여 동일한 영역을 맴도는 경우가 많을 뿐만 아니라, 360도 회전 행동을 주기적으로 수행하기 때문으로 보인다. 또, CoW는 LGX보다는 높지만, ESC와 Ours보다는 낮은 SPL 성능을 보인다. 이는 FBE 알고리즘을 통해 탐험 영역 확장은 용이하지만, 가까운 프론티어부터 순차적으로 방문하여 작업 환경에 따라 이동 효율성이 떨어지기 때문으로 볼 수 있다. 반면, ESC와 Ours에서는 위의 두 모델에 비해 높은 SPL 성능을 보이는데, 이는 PSL을 사용하여 우선순위 탐험을 진행하므로 이동 과정이 효율적이기 때문이라고 할 수 있다. 그러나, ESC 모델 대비 Ours 모델이 1.26배 높은 향상률을 보이며, 이는 제안 모델이 단순히 에이전트 근처의 물체의 위치만을 이용하는 것이 아닌 과거 이력들을 포함하는 의미적 맥락 지도를 활용하여 환경 구성 파악 능력이 뛰어나며, 이로 인해 LLM의 답변의 정확성을 높였기 때문이다. 한편 SR 비교 시에도 Ours 모델이 모든 데이터 집합에서 가장 높은 성공률을 보이는 것을 확인할 수 있다. 특히 작업 성공은 목표 그라운딩 능력이 큰 영향을 미친다. 즉, 앞서 언급한 거짓 양성, 그리고 목표 물체임에도 목표로 탐지해내지 못하는 거짓 음성(false negative)의 경우들을 줄일 때 높은 작업 성공 능력을 보이게 된다. 따라서 모든 데이터 집합에서의 SR 성능을 기반으로 목표 그라운딩 모듈의 우수성을 판단할 수 있다.

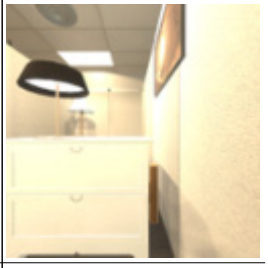
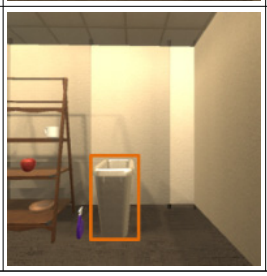
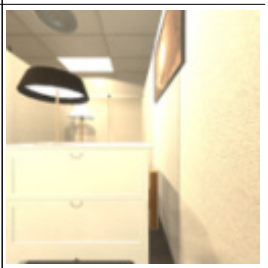
다섯 번째 실험은 PASTURE 데이터 집합에서의 대표적인 탐지 사례들을 중심으로 제안 방식의 성능을 정성적으로 분석해보기 위한 실험이다. [Table 6]은 Uncommon, Spatial (normal), Appearance (normal), Hidden (normal)의 4가지 데이터 집합에 대해 (a) OWL-ViT, (b) GLIP-L, (c) GLEE, (d) GLEE+GLIP (ours) 등 서로 다른 방식의 탐지 결과들을 보여준다. 먼저, [Table 6]의 (a)에서 볼 수 있듯이, 'Whiteboard'가 목표 물체로 주어진 Uncommon 시나리오에서 OWL-ViT는 해당 물체를 잡아내지 못했다. 반면, GLIP은 'Whiteboard'에 해당하는 영역이 아닌, 벽 전체를 잡아내는 것을 통해 해당 물체가 나타내는 의미를 올바르게 이해했다고 말하기는 어렵다. 반면, GLEE와 GLEE+GLIP은 Whiteboard가 위치하는 영역을 정확하게 탐지해내는 것을 알 수 있다. 다음으로 [Table 6]의 (b) Spatial에서 'garbage can on the floor near a spray bottle, bowl, and shelving unit'과 같이 목표 물체를 중심의 공간적 관계가 묘사된 복잡한 문장이 주어진 시나리오에서, OWL-ViT는 목표 물체를 탐지하지 못했으며, GLIP은 문장에 나열된 모든 물체를 탐지하였으나, 핵심 물체인 garbage can에 대한 인식이 부족함을 확인할 수 있다. 그러나 GLEE와 GLEE+GLIP는 공간적 배치를 올바르게 파악하고 해당 위치에 있는 목표 물체만을 정확히 탐지하여 높은 공간 이해 능력을 보여주었다. [Table 6]의 (c) Appearance에서는 'green, potted queen of hearts plant'라는 식물의 색상과 모양을 설명하는 문장이 주어진 시나리오이다. 이때는 특히 형용하는 단어가 앞쪽에 배치되며, 훨씬 어려운 문장이 주어졌을 때도 문장 성분 별 관계를 파악하고 목표 물체의 위치를 판단해낼 수 있어야 한다. Spatial과 비슷하게 OWL-ViT는 물체를 탐지해내지 못하였으며, GLIP은 물체의 경계상자 위치는 올바르게 예측하였지만, plant가 아닌 queen이라는 레이블을 출력하는 것을 통해 문장 내 형용하는 단어들을 올바르게 이해했다고 보기 어렵다. 반면, GLEE와 GLIP의 경우엔 올바르게 목표 물체

[Table 5] Comparison with other SOTA agent models

Agent Model	SR(%) / SPL(%) on PASTURE													
	Uncommon		Spatial				Appearance				Hidden			
	normal		normal		distract		normal		distract		normal		distract	
	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
CoW[1]	31.39	19.98	17.22	9.63	15.28	9.24	28.06	15.47	18.33	11.62	19.17	11.89	15.83	10.22
LGX[2]	16.9	12.1	24.17	15.34	25.55	15.98	24.17	17.68	24.72	16.78	5.83	2.78	5.83	2.78
ESC[3]	32.22	19.54	30.27	15.82	20.00	14.54	30.83	18.90	19.16	13.31	3.06	0.88	3.61	1.27
AML (Ours)	**34.44**	**24.62**	**31.11**	**16.44**	**27.50**	**17.38**	**33.32**	**25.02**	**28.97**	**17.59**	**1.83**	**0.99**	**2.22**	**0.98**

의 올바른 경계상자 위치와 레이블을 출력해내는 것을 확인할 수 있다. 마지막으로, [Table 6]의 (d) Hidden에서는 목표 물체가 존재하지 않고, “alarm clock in the shelving unit drawers”와 같이 해당 물체를숨긴 장소가 문장으로 주어진 시나리오에서 OWL-ViT는 물체를 숨긴 서랍장을 탐지해내는 것을 확인할 수 있다. GLEE는 영상 내 목표 물체의 존재 여부와 상관없이 가장 적합한 물체를 탐지하므로 엉뚱한 물체인 램프를 탐지해내는 것을 확인할 수 있다. 반면, GLEE+GLIP의 경우엔 GLIP에서 목표 탐지를 하지 않아 경계상자 간 IoU값이 0이므로 목표가 없다고 판단하는 것을 확인할 수 있다. 위의 결과들을 토대로, 제안 방식이 다양한 환경과 시나리오에서 목표 물체 그라운딩의 정확성을 향상시킬 수 있음을 확인할 수 있다. 여섯 번째 실험은 PASTURE 데이터 집합에서의 대표 L-ZSON 작업 사례들을 중심으로 제안 방식의 특징과 한계점을 분석해보기 위한 실험이다. [Table 6]은 이 실험의 결과를 나타낸다. 각 그림에서 하향식 지도(top-down view map)의 빨간색 상자와 빨간 핀 모양의 도형은 목표 물체를, 하늘색 상자는 목표 물체를 발견하기 이전까지 탐지한 물체들을, 노란색 상자는 방을 나타낸다. 별 모양의 노란색 도형은 에이전트가 선택한 프론티어 지점을 뜻하며 초록색 핀 모양의 도형은 에이전트의 초기 위치를 나타낸다. 파란색 실선으로 표시된 것은 에이전트의 이동 경로를 나타낸다. 또한, 녹색 상자는 에이전트가 각 위치에서 관측 영상을 나타낸다. [Table 7]의 (a) Uncommon, (b) Spatial, (c) Appearance 시나리오에서는 제안 모델의 에이전트가 목표 물체까지 성공

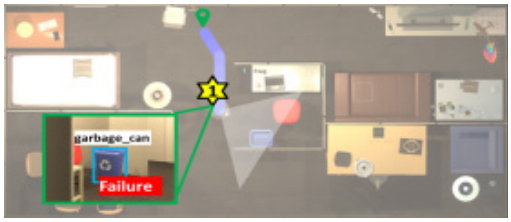
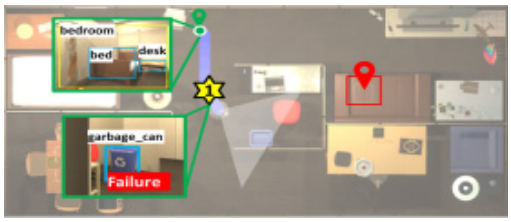
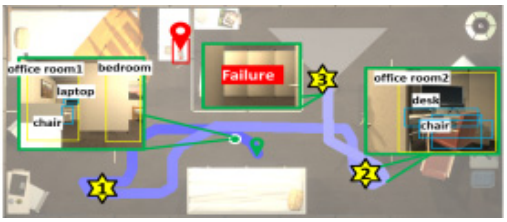
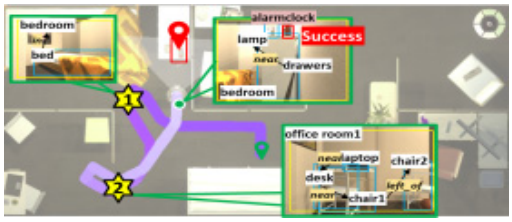
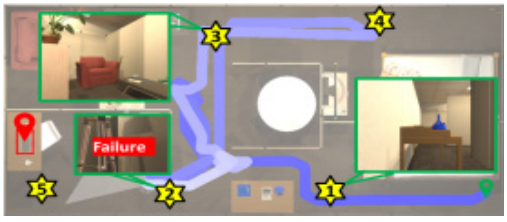
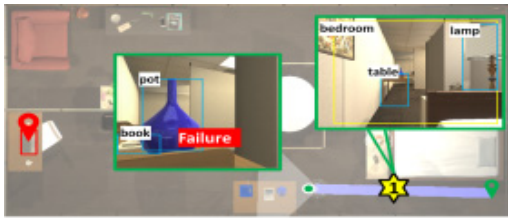
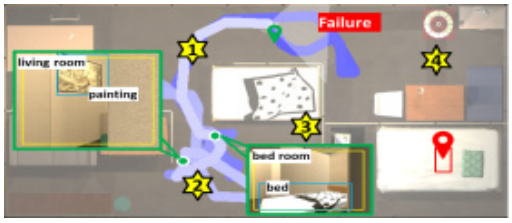
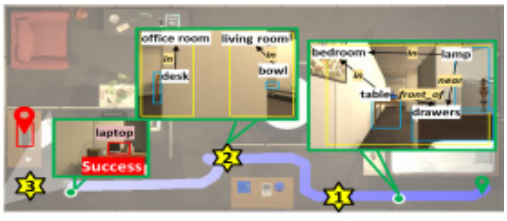
[Table 6] Qualitative comparison between different goal grounding methods

Goal Grounding Method		(a) Uncommon	(b) Spatial	(c) Appearance	(d) Hidden
		“Whiteboard”	“garbage can on the floor near a spray bottle, bowl, shelving unit”	“green, potted queen of hearts plant”	“alarm clock in the shelving unit drawers”
Zero-shot OD	OWL-ViT[1]				
	GLIP[2,3]				
RIS Model	GLEE[9]				
-	GLEE + GLIP				

적으로 도달한 작업 성공 사례이며, (d) Hidden 시나리오에서는 에이전트가 끝내 목표 물체를 찾지 못하고 작업을 실패한 사례를 나타낸다.

[Table 7]의 (a)는 Uncommon 중 우측 상단에 위치하는 목표 물체인 'Crate'를 찾아가야 하는 시나리오이다. 목표 물체인 Crate는 에이전트의 근처의 방에 존재하나, 벽에 의해 가려져 있어 해당 방의 안쪽까지 이동해야만 관측 가능하다. CoW와 ESC에서의 에이전트는 초기 위치에서 360도 회전을 통해 주변

[Table 7] Qualitative analysis of the proposed model

Model	(a) Uncommon	(b) Spatial
	"Crate"	"alarmclock on a dresser near a desk lamp, baseball bat"
CoW		
ESC		
Ours		
Model	(c) Appearance	(d) Hidden
	"grey, plastic laptop"	"apple under the bed with a single pillow"
CoW		
ESC		
Ours		

환경을 탐색하는 과정에서, ‘garbage can’을 관측하자마자 이를 목표 물체로 오인하였고 해당 물체로 바로 이동하여 작업을 실패하였다. 반면, 제안 모델은 회전하면서 ‘garbage can’을 실제 목표인 Crate와 혼동하지 않았다. 또, 회전 과정에서 파악한 환경 구성을 바탕으로 LLM으로부터 living room을 추천받으며, 해당 위치로 이동하여 올바르게 탐지하고 작업을 성공적으로 수행하였다. 이러한 결과를 통해 일반적이지 않은 물체에 대한 언어적 묘사도 정확하게 이해하는 제안 모델의 목표 그라운딩 모듈의 긍정적인 효과를 확인할 수 있었다.

한편, [Table 7]의 (b)는 Appearance 중 좌측 하단에 위치하는 목표 물체인 ‘Laptop’을 찾아가야 하는 시나리오이다. 이때, 목표는 ‘grey, plastic laptop’으로, 색상과 재질을 설명하는 자연어 묘사로 주어진다. 먼저, CoW의 경우엔 초기 위치에서부터 목표 물체와는 전혀 관련없는 프론티어 영역만을 순차적으로 이동하면서 비효율적인 이동 경로를 보여주다, 결국엔 최대 시간 단계인 250 스텝을 초과하여 작업을 실패하였다. ESC는 초기 위치에서 통로로 이동하다 에이전트 시야에 책상 위 존재하는 책을 목표 물체로 오인하여 작업을 실패하였다. 이는 ESC 모델에서는 플라스틱이라는 속성에 대한 언어적 의미를 이해하지 못한 것으로 볼 수 있다. 반면, 제안 모델은 프론티어 중 living room과 office room 중 laptop이 존재할 가능성이 높은 office room을 우선적으로 탐험하면서, 가장 효율적인 경로로 이동하여 작업을 성공하였다.

또, [Table 7]의 (c)는 Spatial 중 알람 시계를 찾아가야하는 시나리오이다. 목표 물체는 ‘alarmclock on a dresser near a desklamp, baseball bat’로 주어지며, 좌측 상단의 침실 근처의 서랍장 위에 놓여있다. 에이전트는 여러 물체가 등장하는 복잡한 문장으로부터 주어를 파악하고, 관련 물체들이 존재할 가능성이 높은 지점을 추론해낼 수 있어야 한다. 먼저, CoW는 초기 위치에서부터 가까운 프론티어 지점을 순차적으로 방문하면서 최대 시간 단계인 250 스텝을 초과하여 작업을 실패하였다. 반면, ESC는 LLM의 답변으로부터 받은 방과 물체 간의 연관성이 떨어져 bedroom이 나왔음에도 관측된 drawers, chair 등을 찾아가면서 작업 실패하였다. 반면, 제안 모델은 LLM으로부터 bedroom과 drawers를 답변으로 받고, 침실 근처를 우선적으로 탐험하면서 목표 물체인 ‘alarm clock’을 방문하여 작업을 성공하였다. 이 결과들을 통해 복잡한 문장을 올바르게 이해하고, 관측 영상으로부터 물체를 찾아내는 에이전트의 목표 그라운딩 모듈의 긍정적 효과를 다시 한 번 확인함과 동시에, PSL을 통한 탐험적 행동 결정 방식의 효율성도 함께 확인할 수 있었다.

마지막으로, [Table 7]의 (d)는 Hidden 중 우측 하단의 침대 아래에 숨겨진 apple을 찾아가야 하는 시나리오이다. 이때, 목표는 ‘apple under the bed with a single pillow’로 주어진다. 먼저, CoW는 에이전트가 초기 탐험하는 과정에서 벽과 벽 사이의 틈에 끼어 이동에 어려움을 겪으며 제자리에서 맴도는 이동이 자주 발생하였고, 결국 최대치인 250시간 단계를 초과하여 작업에 실패하였다. 한편, ESC와 제안 모델에서도 CoW와 유사하게 벽 사이의 틈에 끼어 순환적인 이동 경로를 보였으나, LLM으로부터 ‘bed room’을 답변으로 받아 침실로 이동하면서 이 순환 경로를 벗어날 수 있었다. 그러나 ESC와 제안 모델 모두 경로를 벗어나기 위해 인근 영역을 배회하는 데 많은 시간을 소모하여, 결국 최대 시간 단계인 250 스텝을 초과해 작업에 실패하였다. 추측되는 원인으로는 목표 물체와 유사한 환경 요소를 구분하지 못해 잘못된 지역을 반복적으로 탐색했기 때문으로 보인다. 즉, 해당 환경에서는 중앙에 위치하는 베개가 없는 침대와, 우측 하단에 위치하는 베개가 있는 침대가 존재할 때, 작업 성공을 위해 에이전트는 후자의 침대를 향해 이동하여야 한다. 그러나 에이전트는 이동 작업 초반에 실제 관측한 중앙의 침대 근처의 프론티어들을 추천받게 되어 인근 영역만을 배회하다가 실패한 것으로 보인다. 이와 같은 현재 제안 모델의 한계점을 해결하기 위해서는 보다 정교한 점유 지도 작성이 이루어질 수 있도록 지도 작성 모듈을 개선하거나, 위와 같은 인식 오류를 고려하여 교착상태에 빠지더라도 효과적으로 발견 및 복구하기 위한 기법을 도입하는 등 추가적인 연구가 필요할 것으로 판단된다.

5. 결 론

본 논문에서는 제로-샷 목표 물체 탐색 이동(Zero-shot Goal Object Navigation, L-ZSON) 작업을 위해 시각-언어 모델(VLM)과 대규모 언어 모델(LLM) 등 사전 훈련된 인공지능 기저 모델들을 활용하는 새로운 에이전트 모델 AML을 제안하였다. 제안 모델 AML은 자연어 목표 물체 묘사의 시각적 그라운딩을 위해 참조 영상 분할이 가능한 인공지능 기저 모델인 GLEE[9]를 채용함으로써, L-ZSON 작업에서 요구하는 수준의 자연어 이해 능력과 목표 물체 탐지 능력을 별도의 큰 훈련과정없이 제로-샷 방식으로 활용할 수 있다. 또 제안 모델 AML은 시각 인식 기저모델인 제로-샷 물체 탐지기 GLIP을 이용하여 이동 계획 수립에 필요한 장애물들의 점유 지도뿐만 아니라 환경을 구성하고 있는 방, 가구, 기타 물체들 간의 공간적 배치 관계들도 나타낼 수 있는 의미적 맥락 지도도 생성한다. 또한 제안 모델 AML은 효율적인 탐험적 행동 결정을 위해 대규모 언어 모델 LLM의 상식을 효과적으로 이용하기 위한 맥락정보 기반의 프롬프팅 방법과 LLM의 응답을 토대로 하는 확률적 추론 방법도 이용한다. 그리고 본 논문에서는 3차원 가상 시뮬레이터인 RoboTHOR와 L-ZSON 벤치마크 데이터 집합인 PASTURE를 이용한 다양한 실험들을 통해, 제안 모델의 우수성을 입증하였다.

한편, 앞서 정성적 평가 실험에서 언급한 것과 같이, 현재의 제안 모델은 여전히 주변 환경 물체들과 지속적으로 충돌하면서 특정 지역 내에 갇혀 작업에 실패하는 경우가 가끔씩 나타난다. 추측하는 원인으로는 부정확한 점유 지도로 인해 잘못된 프론티어를 선정하고, 이로 인해 장애물과의 충돌이 발생하였기 때문으로 보인다. 제안 모델의 이러한 현재의 한계점을 해결하기 위해서는 보다 정교한 점유 지도 작성이 이루어질 수 있도록 지도 작성 모듈을 개선하거나, 위와 같은 인식 오류를 고려하여 교착상태에 빠지더라도 효과적으로 발견 및 복구하기 위한 기법을 도입하는 등 추가적인 연구가 필요할 것으로 판단된다.

References

[1] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: baselines and benchmarks for language-driven zero-shot object navigation," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 23171-23181, 2023, DOI: 10.1109/CVPR52729.2023.02219.

[2] V. S. Dorbala, J. F. Mullen, and D. Manocha, "Can an embodied agent find your "Cat-shaped Mug"? LLM-based zero-shot object navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4083-4090, May, 2024, DOI: 10.1109/LRA.2023.3346800.

[3] K. Zhou, K. Zheng, C. Pryor, Y. Shen, H. Jin, L. Getoor, and X. E. Wang, "Esc: exploration with soft commonsense constraints for zero-shot object navigation," *40th International Conference on Machine Learning (ICML)*, Honolulu, Hawaii, USA, pp. 42829-42842, 2023, [Online], https://dl.acm.org/doi/10.5555/3618408.3620214.

[4] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, M. Aravindh, A. Anurag, D. Mostafa, S. Zhuoran, W. Xiao, Z. Xiaohua, K. Thomas, and N. Houlsby, "Simple open-vocabulary object detection," *European Conference on Computer Vision (ECCV)*, Tel Aviv, Israel, pp. 728-755, 2022, DOI: 10.1007/978-3-031-20080-9_42.

[5] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, W. Lijuan, Y. Lu, Z. Lei, H. Jenq-Neng, C. Kai-Wei, and J. Gao, "Grounded language-image pre-training," *2022IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, Louisiana, USA, pp. 10965-10975, 2022, DOI: 10.1109/CVPR52688.2022.01069.

[6] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "Voronav: Voronoi-based zero-shot object navigation with large language model," *arXiv preprint arXiv:2401.02695*, 2024, DOI: 10.48550/arXiv.2401.02695.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, and D. Amodei, "Language models are few-shot learners," *34th International Conference on Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020, [Online], https://dl.acm.org/doi/abs/10.5555/3495724.3495883.

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023, DOI: 10.48550/arXiv.2307.09288.

[9] J. Wu, Y. Jiang, Q. Liu, Z. Yuan, X. Bai, and S. Bai, "General object foundation model for images and videos at scale," *arXiv:2312.09158*, 2024, DOI: 10.48550/arXiv.2312.09158.

[10] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford, L. Weihs, M. Yatskar and A. Farhadi, "RoboTHOR: An open simulation-to-real embodied ai platform," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 3164-3174, 2020, DOI: 10.1109/CVPR42600.2020.00323.

[11] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Lasvegas, NV, USA, pp. 779-788, 2016, DOI: 10.1109/CVPR.2016.91.

[12] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *European Conference on Computer Vision (ECCV)*, pp. 213-229, Nov., 2020, DOI: 10.1007/978-3-030-58452-8_13.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: transformers for image recognition at scale," *2021 International Conference on Learning Representations (ICLR)*, 2021, [Online], https://openreview.net/forum?id=YicbFdNTTy.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 10012-10022, 2021, DOI: 10.1109/ICCV48922.2021.00986.

[15] H. Du, X. Yu, and L. Zheng, "Learning object relation graph and tentative policy for visual navigation," *European Conference on Computer Vision (ECCV)*, pp. 19-34, 2020, DOI: 10.1007/978-3-030-58571-6_2.

[16] S. Zhang, X. Song, Y. Bai, W. Li, Y. Chu, and S. Jiang, "Hierarchical object-to-zone graph for object navigation," *2021IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, pp. 15130-15140, 2021, DOI: 10.1109/ICCV48922.2021.01485.

최 정 현

2022 경기대학교 컴퓨터공학부(학사)
2022~현재 경기대학교 컴퓨터과학과(석사)

관심분야: 인공지능, 로봇지능, 기계 학습

백 호 준

2022 경기대학교 컴퓨터공학부(학사)
2022~현재 경기대학교 컴퓨터과학과(석사)

관심분야: 인공지능, 로봇지능, 컴퓨터 비전

박 찬 솔

2019 광주대학교 국방기술학부(학사)
2019 광주대학교 경찰법행정학부(학사)
2024~현재 경기대학교 컴퓨터과학과(석사)

관심분야: 인공지능, 로봇지능, 기계 학습

김 인 철

1985 서울대학교 수학과(학사)
1987 서울대학교 전산과학과(석사)
1995 서울대학교 전산과학과(이학박사)
1996~현재 경기대학교 컴퓨터공학부 교수

관심분야: 인공지능, 지능로봇, 지식 표현 및 추론